Imagining a Locality Service. (A DRAFT COPY ONLY from the iDigBio GWG). This copy dated 9 Feb 2015. This is a working copy and in no way represents a final version. It’s a compilation of possibilities. It represents the GWG envisioning the possiblities for locality services and how they might be implented.

[Locality Service:](#h.ljjkuxf5bxl9)

[Rationale:](#h.dofdeeqwjy66)

[Contributors:](#h.usq5c5z26pbl)

[Requirements:](#h.x62im13q8fv7)

[Questions:](#h.9nw9mqecalu1)

## Locality Service:

The purpose of this document is to describe the functional requirements of a locality service that provides a mechanism whereby data associated with previously georeferenced locality descriptions may be leveraged in the process of georeferencing.

## Rationale:

Current georeferencing workflows available in the natural history collections community today often result in projects georeferencing the same localities over and over again. Development and deployment of a *locality service* as outlined in this document addresses this time-consuming and expensive issue. Future workflows taking advantage of this service would see several benefits: faster georeferencing as well as improved efficiency and accuracy. In addition, a service of this type facilitates community input (annotation) for georeferences.

## Contributors:

Nelson Rios, Jessica Utrup, Deborah Paul, John Wieczorek, Una Farrell, Reed Beaman, and others in the GWG.

## Requirements:

1. **A locality service should be able to accept and return locations in the most accurate spatial representation available, including metadata about the process used to generate the spatial representation.**
Georeferencing, as applied to natural history collections, is the inference of geospatial geometries from recorded textual locality descriptions. It is common practice to express the geospatial geometry of a collecting event as a single point representing latitude and longitude, often including a measurement of uncertainty in the form of a linear distance from that point. Furthermore, metadata about the process of deducing the georeferenced result is also often recorded. In addition to point representations, collecting events may be represented as multi-points, lines, multi-lines, polygons, multi-polygons, or other geometries. Geohashes and MGRS Raster representations, e.g., probability surfaces are another type of representation (Guo and Wieczorek).
2. **Expose data through an API.**
API to allow querying of data in the locality service. Tools like GEOLocate could integrate this service into its georeferencing capabilities. Some examples:
	1. Give me all records that match (or are similar to) 7.8 miles N of Bogalusa.
		1. This would allow tools like GEOLocate to integrate results from locality service queries into its georeferencing results.
	2. Give me all records collected on or near the Pearl River.
		1. May be difficult to implement and may not be as critical if spatial extent query is implemented.
	3. Give me all records within a defined spatial extent.
		1. This can be very useful to GEOLocate. During georeferencing, an end user could zoom into an area and turn on all “known” past georeferences. A bounding box query would probably be sufficient. A feedback mechanism should be in place, so that inaccuracies or comments about these past records can be submitted back to iDigBio.
	4. Give me all records collected on 11 Dec. 1974 or collected by Royal D. Suttkus
		1. This could also provide useful background data within GEOLocate.
		2. Would collector names need to be standardized?
3. **Allow download of complete datasets.**
Data contained within the locality service should be made available as a downloadable dataset. This would allow for unanticipated use cases and customized analysis.
4. **Data Vetting Process.**
The service would have to deal with variations in georeferencing data quality across numerous sources. Some form of vetting would be required, to ensure we are not just associating inaccurate data from previous georeferencing efforts to existing efforts for the sake of efficiency. Some sources of existing localities may include:
	1. iDigBio/TCNs
	2. Primary data providers
	3. GEOLocate
	4. VertNet et al.
	5. GBIF
	6. Oceanographic Ship Logs
	7. PaleoPortal

Each one of these may require different vetting processes.

1. **Provide secondary data (when available) useful to the georeferencing process.**
Secondary data in this context represents information, not directly describing the physical collection locality, for example, date of collections, collectors, list of taxa etc.
2. **Other Requirements specific to natural history sub disciplines (i.e., paleo & botany)?**
	1. Paleontologic localities also have a temporal component to them.
		1. There are standardized ages for rock units from many countries (GeoLex in the US, British Lexicon in the UK, etc), but historical records contain a host of inaccuracies (ie, the Ordovician Period wasn’t named until 1879. Specimens collected from before that time were placed in either the Cambrian or the Silurian, though they aren’t actually from either of those periods as currently defined).
		2. This temporal component can be useful for narrowing the geographic extent of a locality, so it would be helpful if there was a way to incorporate this as well (such as a locality along the Pearl River in rocks of Pliocene age).

## Versions:

### Version 0: Geography Service (JRW)

Description:

This iteration would be dedicated to basic services at the level of higher geography as defined by the Darwin Core Geography terms.

Data Sources:

Assemble a vetted data set of higher geography data with shape information conforming to at least minimum best practices.

* GADM - standard names and shapes.
* Getty Thesaurus of Geographic Names (TGN) - alternate names.
* VertNet Vocabularies LookupGeography - transformations from non-standard to standard geography.
* Google Maps API

Capabilities (REST) API access to the vetted data set:

* Geography Standardization: Given any combination of Darwin Core geography values (continent, country, countryCode, stateProvince, county, municipality, waterBody, islandGroup, island), return Standardized Darwin Core Geography (SDwCG) values, if they exist, along with the persistent, resolvable GUID (dwc:highergeographyID) and error report (if any). If the combination has never been seen before, add it to the Authority and flag it as unvetted. Queries can be logged and added to the to-vet pool if not already resolved in the Authority.
* Coordinate Standardization: Given lat/lng in any Datum, return lat/lng in WGS84.
* Coordinate Standardization: Given lat/lng in WGS84, return SDwCG values. Start with CRIA’s InfoXY for this?
* Geography Spatial Representation: Given dwc:highergeophyID, return spatial representation of choice (KML, WKT, GeoJSON)
* Return records matching values of geographic terms.
* Download any query result.

Comments:

Establish a vetting protocol that includes independent consensus from multiple vetters.

Secure against misuse, perhaps with authentication.

Determine sustainability model for infrastructure independent of vetting.

Estimate of Effort:

3 months?

Who:

JRW wants to.

### Version 1: Basic Locality Service (JRW)

Description:

This iteration would build upon version 0, adding support to store and retrieve full Darwin Core Location data (all terms in the class dwc:Location - see http://rs.tdwg.org/dwc/terms/index.htm#locationindex). Locations should be immutable, each with a persistent, resolvable GUID (dwc:locationID). This might be the iteration in which to include Darwin Core Event information as well, to get dates and collectors into the mix.

Data Sources:

Assemble a vetted data set of georeferenced Locations with shape information conforming to at least minimum best practices.

* VertNet Georeferences from MaNIS/HerpNET/ORNIS would be a relatively easy start for Locations, a little bit more work for Dates, and a lot more work for Collectors.
* FishNet collaborative georeferences?
* Google Maps API for simple features
* Best practice-compliant Location data from GBIF.

Capabilities (REST) API access to the vetted data set:

* Return Locations intersecting lat/lng (in WGS84)
* Return Locations intersecting geometry
* Return Locations matching values of geographic terms.
* Location Spatial Representation: Given dwc:locationID, return spatial representation of choice (KML, WKT, GeoJSON)

Estimate of Effort:

3 months?

Who:

JRW wants to.

###

### Version 2: Interactive Locality Service (JRW)

Description:

This iteration would build upon version 1, adding support for annotations to Locations and support for replicating Locations to edit. Editable Locations are “unpublished”. Publishing should make a Location immutable because others then have the options to refer to them. available for use by others. Store immutable Locations information with persistent, resolvable GUID (dwc:locationID).

Data Sources:

Assemble additional data sets of georeferenced Locations with shape information conforming to at least minimum best practices as they become available.

* TCNs
* User contributions

Capabilities (REST) API access to the vetted data set:

* Add an editable (unpublished) Location to the service.
* Publishing a Location (making it immutable so that it can be referred to in perpetuity)
* Submit alternate opinion version linked to any Location
* Submit textual annotation to any Location
* Replicate a published Location to an editable unpublished one.

Estimate of Effort:

9 months?

##

## Questions:

What would be the functionality of V0, V1, etc.

How much time will it take (once defined) to get V0 up and running? (What is estimated time-to-deployment)?

How much developer time needed for V0?

Estimated cost for developer time for V0?

What other costs are envisioned?

Who will (optimally) do the development?

What must be in V0 to build the foundation for future versions?

Where will this service live?
What will be required to maintain / sustain it?

Sample “Vetting” process: Confidence scores for georeferences? (or some standard vocabulary for georeferencedBy so someone can “limit to georeferences by “curator” or by “collector”, for example).

(Participation of current collectors - to vet their own locality data).

Suggestion: Integrate existing data, then build combined gazetteer. Build the plumbing - to populate these gazetteers.

As a new proposal - would need to be innovative in resolving the streams of data that exist in the community. (analogous perhaps to FilteredPUSH). Vetting process is the challenge.

Crowd sourcing (“Expert sourcing”) / primary collectors as a resource.

get collectors to vet their locality data.